DOCUMENT RESUME

ED 262 047 TM 850 506

AUTHOR DeMauro, Gerald E.
TITLE Issues in the Placement of Limited-English-Proficient Students and in the Evaluation of Their Instructional Programs.
PUB DATE 1 Apr 85
NOTE 16p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 1-4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Gains; Bilingual Education Programs; Cutting Scores; Elementary Education; English (Second Language); Equivalency Tests; *Evaluation Methods; Grade 3; Grade 4; Grade 5; *Language Tests; *Limited English Speaking; Mathematical Models; Program Effectiveness; Program Evaluation; *Second Language Programs; *Student Placement; Testing Problems; *Test Norms; Test Selection
IDENTIFIERS *Language Assessment Battery; New York (New York)

ABSTRACT

The Language Assessment Battery was administered to third to fifth grade limited English speaking students in New York City. Special techniques addressed two common problems in testing these populations: (1) language tests suitable for limited English speakers change their psychometric characteristics over time as the students' language ability improves; and (2) the absence of a comparison population who do not receive the necessary treatment. A cutting score was determined by the intersection of the target and monolingual curves of scores from an earlier test administration. Students scoring above this point more closely resembled monolingual speakers, while students scoring below this point were more like target students in English abilities. The skills typical of the target range of scoring were compared to those which described the monolingual range of scores; Rasch item calibrations and person abilities were examined. Reading skills (as opposed to listening and writing) best differentiated target and monolingual students. This latent trait model was useful in selecting an appropriate norm. The Chapter I Model A1, using test norms as proxy comparison groups, interprets pretest-posttest differences as growth beyond the improvements made by the norms groups. Suggestions for using this model in limited English speaking programs are made. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED262047

Issues in the Placement of Limited-English-Proficient Students and in the Evaluation of Their Instructional Programs

Gerald E. DeMauro

Kingsborough Community College
of the City University of New York
Brooklyn, New York

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
DeMauro, Gerald E.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL, April 1, 1985

TM 850 506

ABSTRACT

Issues in the Placement of Limited-English-Proficient Students and in the Evaluation of Their Instructional Programs

Identifying Limited-English-Proficient (L.E.P.) students for bilingual education programs must involve assessment of English-language skills. Evaluating bilingual education programs includes measurement of progress in these skills. Since both of these functions require measurement of English-language abilities, it is reasonable that both may be served by a single instrument. On closer examination, one sees that a test which efficiently separates English-proficient from Limited-English-Proficient students for program placement may lack sensitivity to changes in skills among students in the lower-scoring ranges, limiting its validity as an evaluation tool. This issue may be addressed by the choice of norms and the analytic model (DeMauro, 1983b), if the test demonstrates certain psychometric properties, such as homogeneity over the entire scoring range. Possible norms, analytic models, and psychometric criteria are examined with reference to New York City's Language Assessment Battery (LAB-82).

Issues in the Placement of Limited English Proficient[1] Students and in the Evaluation of their Instructional Programs

Gerald E. DeMauro

Programs for Limited English Proficient (L.E.P.) students seriously challenge traditional evaluation models in several ways:

a) They are usually prescribed by law to serve all identified students, thus eliminating comparison populations;

b) The tests have an added burden in maintaining their psychometric properties over time with a population whose native language and English proficiencies are changing as well as their content area skills.

The Chapter I model A1 (Tallmadge and Wood, 1977) addresses the first challenge by using test norms as proxy comparison groups. Students skills are measured in metrics like Normal Curve Equivalent (N.C.E.s) which reference performance to appropriate grade and time of year norms. Differences between pretest and posttest scores, then, may be interpreted as growth beyond the improvements made by the norms groups. [2]

The model assumes that the norms group students and program students were roughly equivalent in the measured skill at the time of pretest. Superior improvements by program students indicate program effectiveness. The model also assumes that the amount of exposure to the subject matter of norms group students and program students is equivalent (DeMauro, 1983a). This assumption allows inference that the rate of growth of the two groups at the time of the pretest is the same, and that the amount the norms group gains in a year constitutes the no-program expectation of the program group.

---

[1] This research was conducted with permission of the New York City Public Schools

[2] Since norms rarely follow students over a school year, the assumption is that students comprising both the spring and fall norms samples represent students in general at those times of year.

This assumption limits the applicability of Model A1 in programs for L.E.P students. For example, tests of English proficiencies, such as the Language Assessment Battery - 82 (Lab-82) (New York City Public Schools, 1982) which provide norms based on monolingual English samples may be used. Pretest scores of bilingual populations, expressed in NCEs could be compared to posttests NCEs. Differences would then be attributed to growth beyond the normal English growth of grade peers.

Unfortunately, the monolingual students who have equivalent pretest scores presumably have had more home exposure (at least) to English than the L.E.P. students. Their English growth rate must be lower [1], limiting the utility of the norms as no program-expecations. Obviously, increases in intensive English experience for the LE.P. group will increase the appropriateness of the model to measure English growth (DeMauro, 1983b).

Several additional evaluation considerations in use of monolingual norms are discussed below. Alternative normative and nonnormative models are offered.

Educational Significance

The attraction of Model A1 is its provision of a means to judge the educational meaning of L.E.P. student improvements. If the programs have the goal of integrating participants into regular monolingual classes, then the monolingual NCEs will show group movement with respect to students in these classes. Alternatives to Model A1 must provide a means of gauging student growth in terms that are educationally meaningful.

[1]Consider rate of growth as the amount gained per unit of English exposure.

Validity Considerations

Many tests of English proficiency, including LAB - 82 serve two purposes, program placement (or population identification) and program evaluation. Necessarily, these two uses are at odds with each other when monolingual norms are employed.

To identify program participants, the instrument must have high taxonomic validity (DeMauro, 1983b). Essentially, it is most valid when it separates L.E.P. and English proficient students into two groups whose scoring ranges have no overlap. With reference to the monolingual norms, then, all, or nearly all, of the L.E.P. sample has achieved an N.C.E. of 1.

For evaluation purposes, however, the instrument must be sensitive to changes in English skills among the L.E.P. students. The more sensitive the test is to growth among L.E.P. students, the less valid it is for placement, and vice versa.

For example, in the perfectly discriminating test, the least proficient L.E.P. student scores a 1, as does the most proficient. If the poorest student makes so much improvement as to become as skilled as the best, use of monolingual norms will not show it.

Conversely, in the perfectly nondiscriminating test the least proficient L.E.P. student again scores a 1, while the most proficient scores a 99. The ranges of the L.E.P. and norms groups completely overlap. While such an instrument is undoubtedly sensitive to skills improvements, it cannot possibly help identify who are L.E.P. students. Any normative evaluation model must, then, address the issue of double use of the test (as must any normative placement model).

Proper Norm

This raises the question of whether the norms used in evaluation need be derived from a monolingual sample. In fact, LAB - 82 has norms available for monolingual and "target", or students being screened for program identification, groups. While these norms offer greater sensitivity to English acquisition, they provide data which have limited educational meaning. For example, achievement is gauged in terms of other L.E.P. students, most of whom are receiving the same program services.

Another feasible normative group are L.E.P. project students delineated by grade level and program experience (DeMauro, 1983b). From year to year, growth may be referenced to proper increasing grade levels and also to a constant experience level. In this way any gains with respect to the norms may be attributed to increased program experience. For example, suppose a third grader is tested in Spring, and placed in a bilingual program. His/her score would be referenced to third graders with no program experience. One year later, the same student could be post tested. This time, however, the score would be referenced to fourth graders with no program experience. The same performance could be used as a pretest for the coming program year by converting the score to an NCE for fourth graders with one year experience. Differences between pretest and posttest scores could not be attributed to maturation, but could be credited to increased program experience.

The preparation of such norms entails testing a very large sample delineated by program experience. Some compromise may be possible by collapsing the high school grades, where the effects of maturation are small (DeMauro, 1983c). One would at least want to sample groups having no years, one year, two years, and three or more years or program experience in each grade group.

Other Psychometric Considerations

Cutoff - The cutoff consideration is not separate from taxonomic validity concerns. The value of each score, as a potential cutoff may be measured by the proportion of L.E.P. students who fail and English proficient students who pass to all students tests (Crehan, 1974). When the chosen score is too low, too many L.E.P. students pass. If it were too high, too many proficient students would fail. In between, an optimum score exists which yields the highest proportion of identified students.

Test validity for evaluation is also affected by the cutoff placement, if monolingual norms are used. In this case, the higher the cutoff, the more students in the program who would score like low skilled monolinguals. Consequently, there would be more room for growth in the program, as measured by these norms.

Reliability. The reliability question is complex. For evaluation purposes, the test must naturally be consistent over time, demonstrating high test - retest stability, or that the test measure the same skills on each administration. This is not such an easy task among a population whose Engish skills are undergoing rapid change.

Homogeneity. Related to this concern is homogeneity. The test should measure the same skills over the whole range of scores. The lowest scoring students should find the same items most difficult as the highest scoring students.

According to Cummins' common underlying proficiency model (Cummins, 1981), English skills of students at all skills levels should be predicted by this underlying linguistic construct. Consequently, scoring varience should be related to underlying proficiency variance.

Moreover, the theory provides the common underlying skill being measured for all levels of English proficiency. We would expect that as students progress from having no English proficiencies to being English proficient, the test items should maintain their relative order, because the nature of acquisition from the start is predicted by their underlying proficiencies.

Other Psychometric Considerations - **continued**

For a test to have any evaluation application, it must compare students at different levels of English skills, on a common trait. This consideration is obvious when one considers comparing pretest (low skill) to posttest (high skill) scores. Such routine comparisons assume that both test results address the same skill.

When the test measures a single skill among all populations, it becomes theoretically possible to mark the progress of a L.E.P. group in terms of a monolingual group. The difficulty may be insufficient overlap in the score distributions to allow use of the norms. In a sense, the monolingual norms raise the test floor.

Blocking Groups by Experience Level

One solution to this floor problem is to group L.E.P. students by experience levels. In this way, severe problems that are most probable among students in their first year of programs may be avoided by students in later years. Data may be analyzed via mixed repeated measures models in which the within subjects variable is pretest to posttest growth and the between subjects variable is experience group (first year, second year, and third or more year students) (DeMauro, 1983b).

This modification of Model A1 is sufficient to chart group effects. However, it may be particularly insensitive to individual growth and/or growth among lower-skilled students. A criterion model is examined below which offers an alternative evaluation system.

Criterion Model

An advantage of a test with L.E.P. and English proficient norms is it permits analysis of the scoring distributions of both populations. For example, one can find the relative percentage of English proficient and L.E.P. students scoring at each raw score of the tests. When one subtracts the L.E.P. percentage from the English proficient percentage, the difference will be positive in the higher scoring ranges and negative in the lower. The point at which the sign of the differences changes is where students are scoring less like English proficient and more like L.E.P. students.

Other Psychometric Considerations - **continued**

For a test to have any evaluation application, it must compare students at different levels of English skills, on a common trait. This consideration is obvious when one considers comparing pretest (low skill) to posttest (high skill) scores. Such routine comparisons assume that both test results address the same skill.

When the test measures a single skill among all populations, it becomes theoretically possible to mark the progress of a L.E.P. group in terms of a monolingual group. The difficulty may be insufficient overlap in the score distributions to allow use of the norms. In a sense, the monolingual norms raise the test floor.

Blocking Groups by Experience Level

One solution to this floor problem is to group L.E.P. students by experience levels. In this way, severe problems that are most probable among students in their first year of programs may be avoided by students in later years. Data may be analyzed via mixed repeated measures models in which the within subjects variable is pretest to posttest growth and the between subjects variable is experience group (first year, second year, and third or more year students) (DeMauro, 1983b).

This modification of Model A1 is sufficient to chart group effects. However, it may be particularly insensitive to individual growth and/or growth among lower-skilled students. A criterion model is examined below which offers an alternative evaluation system.

Criterion Model

An advantage of a test with L.E.P. and English proficient norms is it permits analysis of the scoring distributions of both populations. For example, one can find the relative percentage of English proficient and L.E.P. students scoring at each raw score of the tests. When one subtracts the L.E.P. percentage from the English proficient percentage, the difference will be positive in the higher scoring ranges and negative in the lower. The point at which the sign of the differences changes is where students are scoring less like English proficient and more like L.E.P. students.

Criterion Model - continued

The relative percentage of English proficient students at or above each score plus the relative percentage of L.E.P. students below that score divided by 200 should provide an estimate of the taxonomic validity of the test, using that score as a cutoff (via a modification of Crehan, 1974). Table 1 shows these estimates at different Spring scores of Level 2 of the English LAB - 82, Form A.

Illustrative Study

In the current study, a cutoff score was determined by the intersection of the target curve and the monolingual curve of scores on the Spring administration norms for Level 2 (Grades 3 - 5) Form A of LAB -82. Theoretically, students who score above this point are more like monolingual students in English skills, while students who score below this point are more like target students in English abilities.

The study examines a criterion use of the instrument, in which the skills typical of the target range of scoring are compared to those which describe the monolingual range of scores. To accomplish this, the Rasch item calibrations and person abilities were examined.

Cutoffs established in this way were 58 for third grade, 63 for fourth grade, and 72 for fifth grade. Associated validity estimates (ala Crehan, 1974) were respectively, .730, .680, and .705. Log abilities and standard errors (in parentheses) of these scores were .88 (.25), 1.20 (.26), and 1.88 (.31).

When we examine the items whose calibrations fall within .5 a standard error of the cutoff, it is clear that among the skills of listening, reading, and writing, reading differentiates target and monolingual students.

Insert Table 1 about here.

Table 1

Taxonomic Validity Estimates
at Various Cutoff Scores

Raw Score	Third Grade p[1]	t[2]	b	v	Fourth Grade p	t	b	v	Fifth Grade p	t	b	v
75	57	6	63	.685	51	16	67	.665	33	25	58	.710
74	55	7	62	.690	49	18	67	.665	31	27	60	.700
73	53	8	61	.695	47	19	66	.670	29	30	59	.705
72	51	9	60	.700	45	21	66	.670	27	32	59	.705
71	49	10	59	.705	42	23	65	.675	25	34	59	.705
70	47	11	58	.710	40	24	64	.680	23	35	58	.710
69	45	12	57	.715	39	26	65	.675	20	37	57	.715
68	43	13	56	.720	37	27	64	.680	19	38	57	.715
67	41	14	55	.725	35	29	64	.680	18	40	58	.710
66	39	15	54	.730	33	31	64	.680	17	42	59	.705
65	37	17	54	.730	32	32	64	.680	16	43	59	.705
64	35	18	53	.735	30	34	64	.680	15	45	60	.700
63	34	19	53	.735	29	35	64	.680	14	46	60	.700
62	33	20	53	.735	28	37	65	.675	13	47	60	.700
61	32	21	53	.735	27	38	65	.675	12	48	60	.700
60	31	23	54	.730	26	40	66	.670	11	49	60	.700
59	29	24	53	.735	24	41	65	.675	11	51	62	.690
58	28	26	54	.730	22	43	65	.675	10	53	63	.685
57	27	27	54	.730	20	44	64	.680	10	54	64	.680
56	26	29	55	.725	19	46	65	.675	9	55	64	.680
55	24	30	54	.730	19	49	67	.665	9	57	66	.670

[1] Number of English proficient students below cutoff, per 100 tested.
[2] Number of target students above cutoff per 100 tested.

Illustrative Study - **continued**

In third grade, eight items fell between 0.76 of 1.00 logits. Seven were reading items, four involving use of nouns, two of adjectives, and one of an adverb. The eighth item, a writing item, involved the use of an adjective.

In fourth grade, five items fell between 1.07 and 1.34 logits. Four were reading items, two were verbs (one in the past tense), one noun and one adjectives. The fifth, a writing item, involved use of a past-tense verb.

In fifth grade, only one item fell within the 1.70 to 2.08 logit range. This was a reading item (1.77) using a past tense verb.

Extending the analysis, we examined the items that fell between the median scores of the target norms and the cutoff scores. The target medians in raw scores are shown in Table 2.

TABLE 2

ITEMS BETWEEN TARGET MEDIAN

SCORES AND CUTOFFS

GRADE LEVEL	TARGET MEDIAN	LOGIT MEDIANS	Number of Items From Target Median to Cutoff		
			LISTENING	READING	WRITING
3	44	0.09	3	16	8
4	53	0.59	1	14	3
5	59	0.93	0	9	6

Again, it is clear that among listening, reading, and writing skills, reading is the best discriminator of "target-like" skills from "English proficient-like" skills. In fact, there were 30 listening, 36 reading, and 20 writing items. In grades 3-5 the orders of item difficulties were significantly associated with whether the items were listening, reading, or writing ($\chi^2(4)=50.900, p < .001$, for third grade; $\chi^2(4)=44.820, p < .001$ for fourth grade; and $\chi^2(4)=33.281, p < .001$ for fifth grade).

Possible Analyses

There are several creative ways of analyzing data generated in this model. For example, the cutoff score will increase in the spring of each school year, and will also be higher in each grade level. Distance of an L.E.P. student's score to the cutoff score could be measured, and as time passes, this distance would be controlled for maturation and the English experience of the monolingual sample. Since each cutoff score is associated with a person ability score and a standard error, the distance measured may be made in normalized logits. Pretest distances could be compared to posttest differences.

A more orthodox criterion approach to the educational meaning of the gains would be to decide at pretest which skills were deficient in the students, based upon their ability scores and the item difficulty distribution. Homogenous class groupings may be made, with long term curriculum based on strengthening these identified needs. Curriculum planning should include identifying the skills that should be mastered at different points in the program, as well as contingency plans for when the skills are mastered or when the student fails to master the skills. Such a model may build in some control by basing the projections of where students should be on the cutoff scores appropriate for the grade of the student when the measures will be taken.

SUMMARY AND CONCLUSIONS

Several modifications of the Chapter I Model A1 are suggested for use with L.E.P. students. Care must be taken that the assumptions underlying the model are upheld. These assumptions are explained. The test used should also demonstrate that it is reliable and homogenous over the range of abilities.

Finally, a latent trait model is suggested that has criterion-related interpretations. The model is explored as a means of circumventing the problem with choosing an appropriate norm while still providing the educational significance that is available with normative models.

REFERENCES

Crehan, K.D. Item analysis for teacher-made mastery tests. Journal of Educational Measurement, 1974, 11(4), 225-262

Cummins, J. Four misconceptions about language proficiency in bilingual education. National Association of Bilingual Education Journal, 1981, 5 (3), 31-45.

DeMauro, G.E. Models of bilingual education evaluation. In G. DeGeorge (Ed.) Bilingual Education Journal. Cambridge, Mass., Lesley College, 1983a

DeMauro, G.E. The paradox of the proper norm for ESL evaluation. In S. Seidner (Ed.), Issues of Language Assessment Volume II: Language Assessment and Curriculum Planning, Evanston, Illinois, National College of Education, 1983b, 45-57.

DeMauro, G.E. Grade level and English acquisition: Age, year of birth, or something else? In S. Seidner (Ed.) Issues of Language Assessment Volume III: Language Assessment and Curriculum Planning, Evanston, Illinois, National College of Education, 1983c, 163-170.

Language Assessment Battery - 82 New York: New York City Public Schools, 1982

Tallmadge, G.K. and Wood, C.T. User's Guide: ESEA Title I Evaluation and Reporting System. Mountain View, Ca.: RMC Research Corporation, 1976